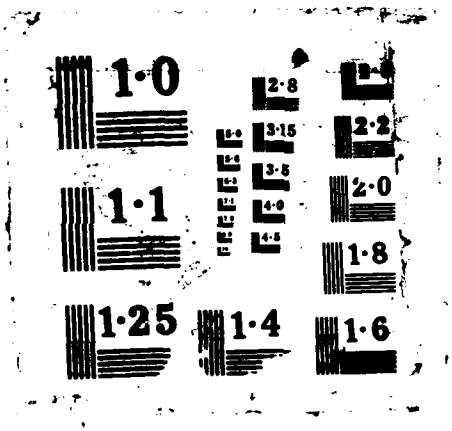
AD-A185 101 QUANTILE STATISTICAL DATA ANALYSIS(U) TEXAS A AND M 1/1
UNIV COLLEGE STATION DEPT OF STATISTICS E PARZEN
FEB 87 TR-Q-1 ARO-23010.1-MA DAAL03-87-K-0003
UNCLASSIFIED F/G 12/3 NL


1·0
1·1
1·25
1·4
1·6
1·8
2·0
2·2
2·8
3·15
3·5
4·0
4·5

ARO 23010.1-MA

TEXAS A&M UNIVERSITY
COLLEGE STATION, TEXAS 77843-3143

Department of
STATISTICS
Phone 409 - 845-3141

QUANTILE STATISTICAL DATA ANALYSIS

Emanuel Parzen

Department of Statistics
Texas A&M University

Technical Report No. Q-1

February 1987

Texas A&M Research Foundation
Project No. 5641

'Functional Statistical Data Analysis and Modeling'

Sponsored by the U. S. Army Research Office

Professor Emanuel Parzen, Principal Investigator

Approved for public release; distribution unlimited

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER ARO 23010.1-MA	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Quantile Statistical Data Analysis		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Emanuel Parzen		8. CONTRACT OR GRANT NUMBER(s) DAAL03-87-K-0003
9. PERFORMING ORGANIZATION NAME AND ADDRESS Texas A&M University Department of Statistics College Station, TX 77843		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS		12. REPORT DATE February 1987
		13. NUMBER OF PAGES
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

quantile functions; data analysis; box plots; Graunt life table.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

This paper presents some reasons why theoretical and sample quantile functions should be routinely used by contemporary statistical data analysts. Quantile methods are introduced in the context of the exponential distribution as a fit to the historically important life table data of Graunt (1661). Section titles are: history of statistics and contemporary textbooks; quantile concepts; identification quantile function; identification quantile box plot; tail classification of probability laws; goodness of fit plots; IQQ plot; cumulative weighted spacings function $D(u)$; quantile simulation and distribution of extreme values; comparison quantile function; nonparametric estimation of probability density; conclusion.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

S/N 0102-LF-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

QUANTILE STATISTICAL DATA ANALYSIS

Emanuel Parzen[1]
Department of Statistics
Texas A&M University

Abstract.

This paper presents some reasons why theoretical and sample quantile functions should be routinely used by contemporary statistical data analysts. Quantile methods are introduced in the context of the exponential distribution as a fit to the historically important life table data of Graunt (1661). Section titles are: history of statistics and contemporary textbooks; quantile concepts; identification quantile function; identification quantile box plot; tail classification of probability laws; goodness of fit plots; IQQ plot; cumulative weighted spacings function $D(u)$; quantile simulation and distribution of extreme values; comparison quantile function; nonparametric estimation of probability density; conclusion.

1. History of Statistics and Contemporary Textbooks.

A central problem of statistical data analysis [that was formulated by 19th century pioneers such as Quetelet (1796-1874) and Galton (1822-1911)] is identifying distributions that fit the data. In *The History of Statistics*, Stigler (1986) writes (p. 268) that these pioneers emphasised the use of normal curves to fit data; they 'proposed that the conformity of the data to this characteristic [normal] curve was to be a sort of test of the appropriateness of classifying the data together in one group; or rather the nonappearance of this curve was indicative that the data should not be treated together.'

By 1875 Galton 'had devised a different way of displaying the data. He ordered the data in increasing order and, effectively, graphed the data values versus the ranks.' Galton used the name 'ogive' for the theoretical form of this curve for a normal distribution; Stigler writes 'we now call it the inverse normal cumulative distribution function'. I call this ideal graph a quantile function of the normal distribution; the graph of ordered data values, denoted $X(j;n)$, versus $(j-.5)/n$ or $j/(n+1)$, is called the sample quantile function, denoted $Q^{\hat{}}(u), 0 < u < 1$.

This paper presents some reasons why theoretical and sample quantile functions should be routinely used by contemporary statistical data analysts. They can be used to not only test the fit (or lack of fit) of a normal distribution to data, but also to describe other general families of distributions and to identify which distributions fit the data.

Textbooks with titles such as *Introduction to Contemporary Statistical Methods* omit many important topics that are actually useful in the theory and practice of statistical data analysis. On my list of important topics (for which I always look in the index and usually fail to find) are: uniform distribution, exponential distribution, order statistics, extreme values, quantile function. Traditional introductory textbooks describe methods based on mean and variance. To qualify as 'contemporary' a textbook adds the following topics: box plot, fences, stem and leaf plot, trimmed and Winsorised sample. In my opinion quantile function interpretations are needed for these topics to acquire beauty and utility that will excite students; however how to do this is not explicitly discussed in this paper.

We introduce the ideas of quantile-based statistical data modeling in the context of the exponential distribution. Let X be a continuous random variable with distribution function $F(x) = \Pr[X \leq x]$ and probability density function $f(x) = F'(x)$.

[1]Research Sponsored by the U. S. Army Research Office Project DAAL03-87-K-0003.

We call $F(x)$ an exponential distribution with parameter λ if

$$1 - F(x) = \exp(-\lambda x), x > 0, f(x) = \lambda \exp(-\lambda x), x > 0$$

Its mean μ equals $1/\lambda$, since (for a non-negative random variable)

$$\mu = \int_0^\infty x f(x) dx = \int_0^\infty (1 - F(x)) dx = \int_0^\infty \exp(-\lambda x) dx.$$

The standard exponential distribution is the exponential distribution with mean 1.

2. Quantile Concepts.

The QUANTILE FUNCTION $Q(u), 0 < u < 1$, is the inverse $x = F^{-1}(u)$ of the distribution function $u = F(x)$. To find $x = Q(u)$ one solves $u = F(x)$.

For an exponential distribution, one obtains $x = Q(u)$ by solving $1 - u = \exp(-\lambda x)$; therefore

$$Q(u) = (1/\lambda) \log(1 - u)^{-1} = \mu(-\log(1 - u))$$

The mean μ of a distribution F or random variable X can be computed from the quantile function Q:

$$\mu = \int_0^1 Q(u) du.$$

The MEDIAN and QUARTILES of a distribution F or random variable X are defined to be

$$Q(.5), Q(.25), Q(.75),$$

the values of $Q(u)$ at $u = .5, .25, .75$. We define QUARTILE DEVIATION DQ by $DQ = 2(Q(.75) - Q(.25))$.

For an exponential distribution, $Q(.5) = \mu \log 2 = .69\mu$; $Q(.25) = \mu \log(4/3) = .29\mu$; $Q(.75) = \mu \log 4 = 1.39\mu$. The interquartile range $Q(.75) - Q(.25) = 1.1\mu$; quartile deviation $DQ = 2(Q(.75) - Q(.25)) = 2.2\mu$.

Two important quantile concepts are $q(u) = Q'(u)$, QUANTILE DENSITY FUNCTION, and $fQ(u) = f(Q(u))$, DENSITY QUANTILE FUNCTION. For F continuous, $F(Q(u)) = u$ and $fQ(u)q(u) = 1$. For a standard exponential distribution, $fQ(u) = 1 - u$.

Two important universal measures of scale of a distribution are DQ and $1/f(median) = 1/fQ(.5) = q(.5)$. They approximately equal each other because DQ is a numerical derivative of $Q(u)$ at $u = .5$.

How do we apply these concepts to determine distributions that fit data? Given data (sample) compute a sample quantile function denoted $Q^\sim(u)$. The sample distribution function is defined by $F^\sim(x)$ =fraction of sample$\leq x$; the sample quantile function $Q^\sim(u)$ is the inverse of $F^\sim(u)$. In terms of the order statistics $X(1;n) \leq \ldots \leq X(n;n)$ of a sample

$$Q^\sim(u) = X(j;n) \text{for} (j-1)/n < u \leq j/n.$$

One usually adopts a continuous version of the sample quantile function defined by linear interpolation between its values

$$Q^\sim((j - .5)/n) = X(j;n), j = 1, \ldots, n.$$

When true mean $\mu = 18$, and the distribution is exponential, $Q(.5) = 12.4, Q(.25) = 5.2, Q(.75) = 25$. If similar values hold for the sample analogues of population parameters (denoted by adding a tilde ($\sim$) to the population notation) one suspects, and conjectures, that an exponential distribution fits.

Table 1. GRAUNT'S LIFE TABLE (1661). OBSERVED PROPORTION AND CUMULATIVE PROPORTION IN VARIOUS INTERVALS OF OBSERVED VALUES OF AGE AT TIME OF DEATH (IN LONDON 1534).

Index j	Age Interval $Q^-(u(j-1)) - Q^-(u(j))$	Proportion $p(j)$	Cumulative proportion $u(j)$
1	0 – 6	.36	$.36 = F^-(6)$
2	6 – 16	.24	$.60 = F^-(16)$
3	16 – 26	.15	$.75 = F^-(26)$
4	26 – 36	.09	$.84 = F^-(36)$
5	36 – 46	.06	$.90 = F^-(46)$
6	46 – 56	.04	$.94 = F^-(56)$
7	56 – 66	.03	$.97 = F^-(66)$
8	66 – 76	.02	$.99 = F^-(76)$
9	76 – 86	.01	$1.00 = F^-(86)$

Table 2. GRAUNT'S LIFE TABLE SAMPLE QUANTILE FUNCTION.

j	0	1	2	3	4	5	6	7	8	9=k
$u(j)$	0.	.36	.60	.75	.84	.90	.94	.97	.99	1.00
$Q^-(u(j))$	0	6	16	26	36	46	56	66	76	86

For an illustrative example we consider Graunt's Life Table data (that should be familiar to all students of statistics). It was published in 1661 by John Graunt, in an attempt to analyse data dealing with age at time of death in London. The original data was collected by Thomas Cromwell in 1534 from Church of England records of births and deaths. Graunt is credited with starting modern statistics by creating Table 1. Brilliant lectures by James R. Thompson of Rice University brought this important data set to my attention.

From Graunt's life table (Table 1) one computes sample mean $\mu^- = 18.22$ (in words, the average age at death was approximately 18 years), $Q^-(.25) = 4.2, Q^-(.5) = 11.8$ (median age at death was approximately 12 years), $Q^-(.75) = 26, DQ = 43.6$. These are found by interpolating the values of the sample quantile function in Table 2.

To compute sample mean (from grouped data) we use formulas

$$\mu^- = \sum_{j=1}^{k} .5(Q^-(u(j-1)) + Q^-(u(j)))(u(j) - u(j-1))$$

$$= \sum_{j=1}^{k} (Q^-(u(j)) - Q^-(u(j-1))(1 - .5(u(j-1) + u(j)))$$

The second formula can be interpreted using the fact that $1 - u$ is the standard exponential density quantile.

It does not seem to be customary in the literature to discuss which distributions fit the data that one is analysing (here Graunt's life table). Techniques are discussed in this paper which can guide the statistical data analyst to identify and test standard parametric distributions (such as the exponential distribution) as a smooth distribution that fits the sample. We discuss the respective roles: (i) $F^-(x)$, sample distribution function, (ii) $Q^-(u)$, sample quantile function, (iii) $F^-(x)$, smooth distribution estimated from data (for Graunt life table, an exponential distribution with mean 18.22), (iv) $Q^-(u)$, smooth quantile function, (v) $D^-(u) = F^-(Q^-(u))$, comparison quantile function, (vi) $D^-(u)$, cumulative weighted spacings, tests constancy of ratio of derivatives $Q^{-\prime}(u)/Q^{-\prime}(u)$, (vii) $QI(u)$, identification quantile function. The statistician's problem

is to develop a framework which explains how and why to use these functions to develop graphical and numerical diagnostics which guide us to identify distributions (such as the normal or exponential) that fit the data.

3. Identification Quantile Function.

The median, which we henceforth denote $MQ = Q(.5)$, is a universal measure of location. It is superior to the mean by the criterion of being more robust (resistant to outliers in the data whose presence will in fact be detected by the identification quantile function). But we recommend the median not because of its robustness but because it forms one of the tools of quantile based methods of statistical data analysis.

Statisticians who favor (or at least teach) mean and standard deviation as measures of location and scale use them to standardize the data by subtracting the mean and dividing by the standard deviation. The quantile based analogy to standardization is to transform the random variable X to

$$XI = (X - MQ)/DQ$$

whose quantile function is

$$QI(u) = (Q(u) - MQ)/DQ$$

We call $QI(u)$ the Identification Quantile Function. Our motivation for introducing this function is that it is approximately equal to the unitized quantile function

$$Q1(u) = (Q(u) - MQ)/Q'(.5) = fQ(.5)(Q(u) - MQ).$$

which has value 0 and slope 1 at $u = .5$. The probability density $f(x)$ corresponding to the unitized quantile function has been normalized so that $f(\text{median}) = 1$. The unitized normal probability density is $f(x) = \exp(-\pi x^2)$.

Universal measures of location and scale are MQ and DQ. Diagnostic measures of skewness are

$$QI(.25), QI(.75), QIM = .5(QI(.25) + QI(.75)), -.25/QI(.25), .25/QI(.75);$$

note that always $QI(.75) - QI(.25) = .5$. Diagnostic measures of (left and right) tail behavior are $QI(.01)$ and $QI(.99)$. A combined measure of tail behavior (useful for probability density estimation) is $QI(.99) - QI(.01)$, called the identification quantile range.

4. Identification Quantile Box Plot.

An identification quantile box plot is a plot consisting of a box from $QI(.25)$ to $QI(.75)$ with a midline at $QI(.5) = 0$ and a cross at QIM. Fences are defined to be $\max(-1, QI(0))$ and $\min(1, QI(1))$. Lines are drawn from identification quartiles to fences. Data values outside the fences are considered outliers or out-and- outliers, depending on whether they are interpreted as representing long tails or blunders. One also indicates the location of (sample mean-MQ)/DQ. The values of identification quartiles and fences are recorded on the plot.

5. Tail Classification of Probability Laws.

Representations of the density quantile function behavior as u tends to 0 or 1 is used to provide a quantitative index of tail behavior which we call the tail exponent. It is used to qualitatively classify tail behavior in three types, called short, medium, and long. Medium tails are further classified in three groups: medium-short, medium-medium, medium-long; a good summary of these concepts introduced by Parsen (1979) is given by Schuster (1984).

These five groups reduce to three groups (short, medium, long) when expressed in terms of hazard rate functions (decreasing, constant, increasing). The right and left hazard functions are respectively defined by

$$h_1(x) = f(x)/(1 - F(x)), h_0(x) = f(x)/F(x).$$

The right and left hasard quantile functions are defined

$$h_1Q(u) = fQ(u)/(1-u), h_0Q(u) = fQ(u)/u.$$

Our classifications of tail behavior can be empirically related to the behavior of the identification quantile function as u tends to 0 or 1. The left tail is classified: $0 > QI(.01) > -.5$, short tail; $-.5 > QI(u) > -1$, medium-short; $-1 > QI(u)$, medium-long and long tail. The right tail is classified short, medium short, or long according as $QI(.99) < .5, .5 < QI(.99) < 1, 1 < QI(.99)$.

For Graunt's Life Table, $QI(.25) = -.17, QI(.75) = .33, QIM = .5QI(.75) + QI(.25) = .08, QI(.01) = -.27, QI(.99) = 1.47$. Experience with typical values of these diagnostic measures for various standard frequently encountered distributions leads one to conjecture that the sample distribution function $F^{\sim}(x)$ of the data in Table 1 is fit by an exponential distribution $F^{\wedge}(x)$ with a suitable estimated mean $\mu^{\sim}$.

6. Goodness of Fit Plots.

To evaluate the fit of a model described by $F^{\wedge}(x)$ or $Q^{\wedge}(u)$ to data described by $F^{\sim}(x)$ or $Q^{\sim}(u)$ one has a bewildering number of options. The theory of goodness of fit tests is concerned with the theoretical study of the many test statistics available, and offers little practical guidance on which methods to use in practice. This extensive literature can only be briefly illustrated in this paper, with emphasis on graphical comparisons.

One can compare plots: (1) $F^{\wedge}(x)$ and $F^{\sim}(x)$ vs. x, on the same graph; (2) $Q^{\sim}(u)$ vs. $Q^{\wedge}(u)$, called Q-Q plot; (3) $D^{\sim}(u) = F^{\wedge}(Q^{\sim}(u))$ vs. u, called D-uniform plot (it is equivalent to a plot of $F^{\wedge}(x)$ vs. $F^{\sim}(x)$ called a P-P plot). We recommend variants of the last method. One can interpret $D^{\sim}(u)$ as sample quantile function of the transformed random variable $U^{\wedge} = F^{\wedge}(X)$. The goodness of fit problem is transformed to tests of fit of $U^{\wedge}$ by a uniform [0,1] distribution and by estimation of the true quantile function, denoted $D(u)$, of $U^{\wedge}$. We call $D^{\sim}(u)$, $0 < u < 1$, a sample comparison quantile function.

When $F^{\wedge}$ is exponential, $D^{\sim}(u) = 1 - \exp(-Q^{\sim}(u)/\mu^{\wedge})$. Its values for Graunt's life data is given in Table 3. Figure 2 presents a IQQ plot as a test of fit of Graunt's life table by an exponential distribution. Figures 3-6 present plots on same graph of sample and smooth distributions. The combinations are $F^{\sim}(x)$ and $F^{\wedge}(x)$ vs. x (Figure 3), $Q^{\sim}(u)$ and $Q^{\wedge}(u)$ vs. u (Figure 4), $Q^{\sim}(u)$ vs. $Q^{\wedge}(u)$, a Q-Q plot (Figure 5), and $F^{\wedge}(x)$ vs $F^{\sim}(x)$, a P-P plot (Figure 6) which also plots $D^{\sim}(u) = F^{\wedge}(Q^{\sim}(u))$. Figures 6 and 7 present $D(u)$ plots as tests of fit of Gaunt's life table by an exponential distribution; $D^{\sim}(u)$ = cumulative weighted spacings in Figure 7.

7. IQQ (Identification quantile - quantile) Plot.

To test whether a sample is normal or exponential, one tests the hypothesis $Q(u) = \mu + \sigma Q_0(u)$ by a scatter plot of $(Q_0(u(j)), Q^{\sim}(u(j)))$ at suitable values $u(j), j = 1, \ldots, k$, in the interval $0 < u < 1$. This plot, called a Q-Q plot, is judged visually for linearity.

We prefer to use what we call a IQQ plot; it is a scatter diagram of $(Q_0I(u(j)), Q^{\sim}I(u(j)))$ with a grid of lines which may make it easier to judge visually for linearity. A IQQ plot for Graunt's life table is given in Figure 2.

8. Cumulative Weighted Spacings Function $D(u)$.

Users of QQ and IQQ plots report that they are difficult to interpret. I propose that one should prefer plots that are graphs of functions such as various functions $D(u), 0 < u < 1$, which can be defined to measure the 'distance' between two distributions.

To compare $Q(u)$ with $\mu + \sigma Q_0(u)$ we recommend comparing their derivatives (equal to $q(u)$ and $\sigma q_0(u)$ respectively). Since σ is unknown we test for constancy the ratio $q(u)/q_0(u) = q(u)f_0Q_0(u)$; equivalently test the deviation from 1 of

$$d(u) = q(u)f_0Q_0(u)/\sigma_0,$$

$$\sigma_0 = \int_0^1 q(t) f_0 Q_0(t) dt.$$

We call $d(u)$ a weighted spacings function, since spacings $X(k;n) - X(k-1;n)$ are the building blocks of estimators of $q(u)$.

One approach to testing $d(u)$ is to estimate and test the deviation (from the uniform function $D_0(u) = u$) of the cumulative weighted spacings function

$$D(u) = \int_0^u d(t) dt$$

The sample analogue of $d(u)$ and $D(u)$ to test exponentiality is: for $u(j-1) < u < u(j)$, $d^{\sim}(u) = d^{\sim}(j)$,

$$d^{\sim}(j) = (Q^{\sim}(u(j)) - Q^{\sim}(u(j-1)))(1 - .5(u(j-1) + u(j)))/\mu^{\sim};$$

$D^{\sim}(u)$ linearly interpolates its values $D^{\sim}(u(j)) = d^{\sim}(1) + \ldots + d^{\sim}(j)$. Note that $\sigma_0{}^{\sim} = \mu^{\sim}$.

Table 3. GRAUNT'S LIFE TABLE $Q^{\wedge}$, $Q^{\sim}$, $F^{\sim}$, $F^{\wedge}(Q^{\sim}) = D^{\sim}$ FOR FITTED EXPONENTIAL $F^{\wedge}(x) = 1 - \exp(-x/\mu^{\sim}), \mu^{\sim} = 18.2$, $D^{\sim}(u)$ CUMULATIVE EXPONENTIAL WEIGHT SPACINGS (CUMWTSPAC).

j	$Q^{\wedge}(u(j))$	$Q^{\sim}(u(j))$	$F^{\sim}Q^{\sim}(u(j))$	$F^{\wedge}Q^{\sim}(u(j))$	$D^{\sim}(u(j))$CUMWTSPAC
0	.09	0	.00	.00	.00
1	8.13	6	.36	.28	.27
2	16.69	16	.60	.58	.56
3	25.26	26	.75	.76	.73
4	33.39	36	.84	.86	.85
5	41.95	46	.90	.91	.92
6	51.26	56	.94	.95	.96
7	63.89	66	.97	.97	.986
8	83.91	76	.99	.98	.997
9	96.54	86	1.00	.99	1.00

Figures 6 and 7 show how we plot $D^{\sim}(u)$ for comparison with $D_0(u) = u$. In addition to the graphical diagnostic of the plot, there are many numerical diagnostics that can be performed.

9. Quantile Simulation and Distribution of Extreme Values.

A general distribution function $F(x), -\infty < x < \infty$, is a non- decreasing function continuous from the right. Its quantile function (or inverse distribution function), defined by

$$Q(u) = \inf\{x : F(x) \geq u\},$$

is a non-decreasing function continuous from the left. It is an inverse under inequality; for any x and u

$$F(x) \geq u \text{ if and only if } x \geq Q(u).$$

An important property of quantile functions is a formula for functions of random variables. THEOREM. Assume g is non-decreasing and continuous from the left. Then $Y = g(X)$ has quantile function

$$Q_Y(u) = g(Q_X(u)).$$

One can represent X in terms of a uniform [0,1] random variable U by $X = Q(U)$ since $Q(U)$ has quantile function $Q(Q_U(u)) = Q(u)$.

When F is continuous, one can transform X to U, a uniform [0,1] random variable, by $U = F(X)$ since $F(X)$ has quantile function $F(Q(u)) = u$.

A random sample $X(1), \ldots, X(n)$ of X can be simulated by generating a random sample $U(1), \ldots, U(n)$ of U, and forming $X(j) = Q(U(j))$. This process, illustrated in Figure 8 for the normal and Cauchy distributions, demonstrates that the quantile function provides a powerful graphical representation of a distribution because of the following equivalence: (1) a random sample of X, (2) observing Q(u), quantile function of X, at a random sample of points on the unit interval. To compare two distributions, such as the normal or Cauchy, one way is to plot (as in Figure 8) graphs of their identification quantile functions plotted on the same scale (the longer tailed one will have to be truncated at a suitable value).

The representation of X in terms of U by $X = Q(U)$ provides a quantile approach to the distribution theory of order statistics and extreme values. Let $X(1;n) < \ldots < X(n;n)$ be the order statistics of a random sample $X(1), \ldots, X(n)$. The kth order statistic $X(k;n)$ has the same distribution as $Q(U(k;n))$ where $U(k;n)$ is the kth order statistic of a random sample from uniform [0,1].

10. Comparison Quantile Function.

A quantile based concept that unifies parameter estimation and goodness of fit hypothesis testing procedures is the comparison quantile function $D(u) = F(G^{-1}(u))$ which compares two distribution functions $F(x)$ and $G(x)$. The comparison quantile density is

$$d(u) = D'(u) = f(G^{-1}(u))/g(G^{-1}(u))$$

The Kullback information divergence can be evaluated by

$$I(G;F) = -\int_{-\infty}^{\infty} (\log(f(x)/g(x))g(x)dx = \int_0^1 -\log d(u)du$$

The graph of $d(u)$ provides insight into the rejection method of simulation. One seeks to generate a sample $X(1), \ldots, X(m)$ from F as an acceptable subset of a sample $Y(1), \ldots, Y(n)$ from $G(x)$. THEOREM. Assume that $D(0) = 0$ and there is a constant c such that $d(u) \leq c$ for all u. Generate two independent uniform [0,1] random variables $U(1)$ and $U(2)$. Acceptance and rejection rule: If

$$U(2) \leq d(U(1))/c,$$

then accept $Y = G^{-1}(U(1))$ as an observed value of X. Otherwise reject Y. (Continue by generating two more uniform [0,1] random variables). The probability of acceptance is 1/c.

The relation between two distributions F and G is best understood by a plot of $u_2 = d(u_1)$.

This plot can be used to graphically describe the rejection rule of simulation and to prove it. Verify that the area under the curve from $u_1 = 0$ to $u_1 = G(x)$ equals $D(G(x)) = F(x)$; the event that $U(1) \leq G(x)$ and $U(2) \leq d(U(1))/c$ has probability $F(x)/c$; the event that $X \leq x$ can be shown to have probability $F(x)$.

11. Nonparametric Estimation of Probability Density.

To identify distributions that fit data, one can use parametric models such as the location-scale parameter model $Q(u) = \mu + \sigma Q_0(u)$, or one can nonparametrically form estimators $f^{\wedge}(x)$ of the probability density function (see Silverman (1986)). We consider only the kernel estimator

$$f^{\wedge}(x) = (1/n)\sum_{j=1}^{n}(1/h)K((x - X(j))/h)$$

where $K(x)$ is a probability density function and h is a bandwidth to be selected.

For K we recommend (Parsen (1962)) the 'Parsen window' which is the probability density of the sum of four uniforms

$$K(x) = \begin{cases} (4/3) - 8x^2 + 8x^3, & 0 < x < .5 \\ (8/3)(1-x)^3, & .5 < x < 1 \\ 0, & 1 < x \\ K(-x), & x < 0 \end{cases}$$

As a first choice to consider for h, by adapting Silverman (1986), p. 47, we recommend

$$hopt = K(0)DQn^{-.2}$$

To accept or reject the goodness of the value of h chosen we judge the deviation from uniformity of the comparison quantile function $D^-(u) = F^(Q^-(u))$. We evaluate this function at $u = (j-.5)/n$ by $F^(X(j;n))$. Other choices of hopt are multiples of hopt based on diagnostics of the tail behavior of the distribution, given by $QI(.99) - QI(.01)$. The deviation of $D^(u)$ from uniformity is used to guide the search for the best value of h for the data being analysed.

The details of this procedure for choosing a kernel probability density estimator cannot be given in this paper. It is best explained by examples of the quality of nonparametric probability density estimators to which it leads for famous data sets (Buffalo snowfall, Yellowstone geyser eruption times) which are used as test cases for density estimation methods (compare Silverman (1986)).

12. Conclusion.

The process of analysing a univariate sample can be viewed as fitting a smooth distribution $F^(x)$ to a sample distribution $F^-(x)$. The process of comparing $F^$ and F^- requires a knowledge of the theory and practice of quantile functions. *'In order to get to the fruit of the tree you have to go out on a limb'* is a proverb that statisticians may take as an omen that they should explore the quantile limb which is always lurking.

References

GRAUNT, JOHN (1665) *Natural and Political Observations Mentioned in a Following Index, and Made upon the Bills of Mortality*, 3rd. ed. London: John Martyn and James Allestry (1st ed. 1662).

PARZEN, EMANUEL (1962) On Estimation of a Probability Density Function and Mode, *Ann. Math. Stat.*, 33, 1065-1076.

PARZEN, EMANUEL (1979) Nonparametric Statitical Data Modeling, *J. Amer. Statist. Assoc.*, 74, 105-131.

SCHUSTER, EUGENE F. (1984) Classification of Probability Laws by Tail Behavior, *J. Amer. Statist. Assoc.*, 79, 936-940.

SILVERMAN, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

STIGLER, STEPHEN M. (1986) *The History of Statistics*, Cambridge: Harvard University Press.

Titles of Figures

Figure 1. Identification quantile plot of Graunt life table.

Figure 2. Identification quantile-quantile (IQQ) plot of Graunt life table vs. exponential distribution.

Figure 3. $F^{\sim}(x)$, Graunt life table sample distribution (dot) and $F^{\wedge}(x)$, exponential mean 18.22 distribution (solid).

Figure 4. $Q^{\sim}(u)$, Graunt life table sample quantile (dot), and $Q^{\wedge}(u)$, exponential mean 18.22 quantile (solid).

Figure 5. Q-Q plot of $Q^{\sim}(u)$ vs $Q^{\wedge}(u)$.

Figure 6. P-P plot of $F^{\wedge}(x)$ vs $F^{\sim}(x)$, same as plot of $D^{\sim}(u) = F^{\wedge}(Q^{\sim}(u))$, $D_0(u) = u$ is also plotted (dots).

Figure 7. $D^{\sim}(u)$, cumulative exponential weight spacings (solid); $D_0(u) = u$ (dot).

Figure 8. Random sample from normal (top) and Cauchy (bottom) represented as values of quantile function $Q(u)$ at random sample from uniform [0,1].

Figure 9. Comparison quantile density $d(u) = D'(u)$, $D(u) = GF^{-1}(u)$, F normal, G Cauchy, $d(u)$ bounded (top), F Cauchy, G normal $d(u)$ unbounded (below).

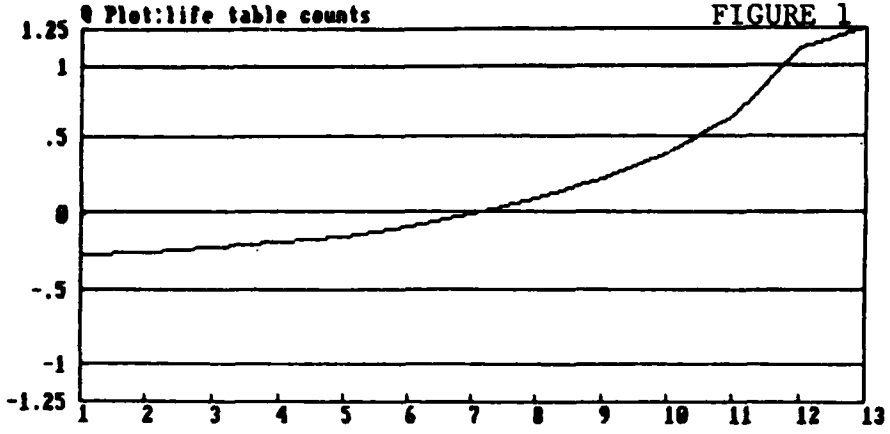


FIGURE 1

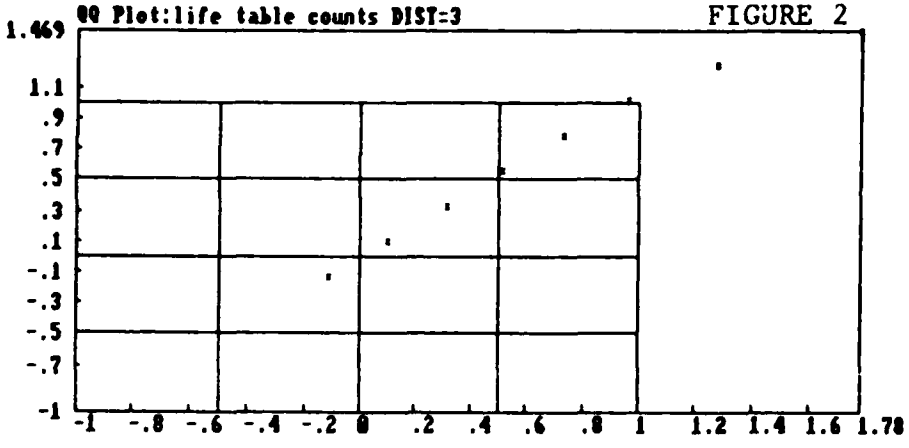


FIGURE 2

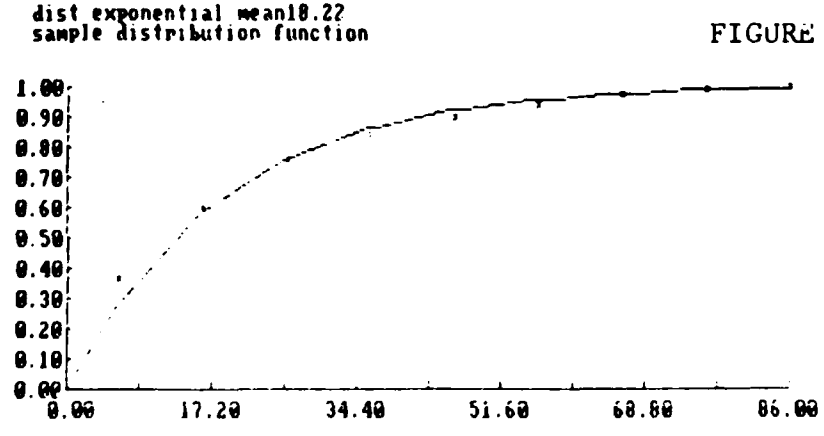


FIGURE 3

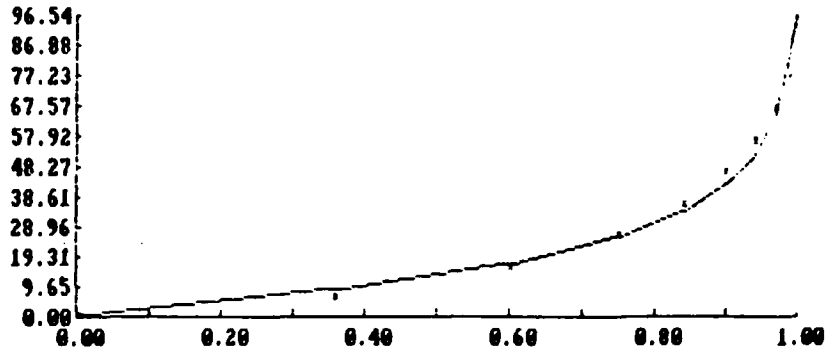


FIGURE 4

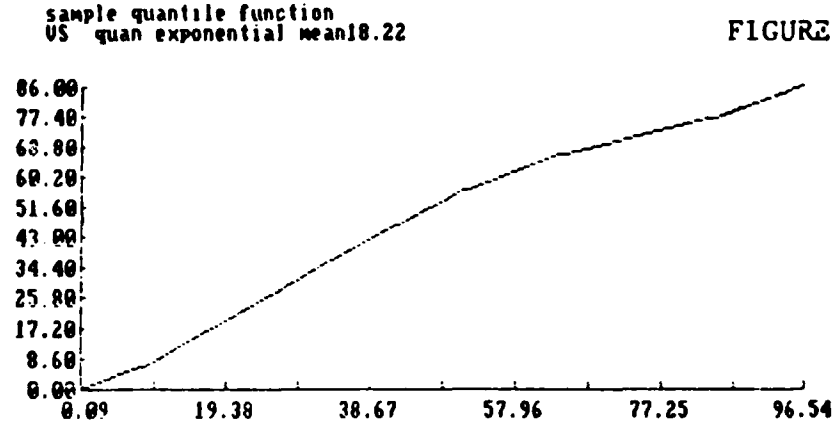


FIGURE 5

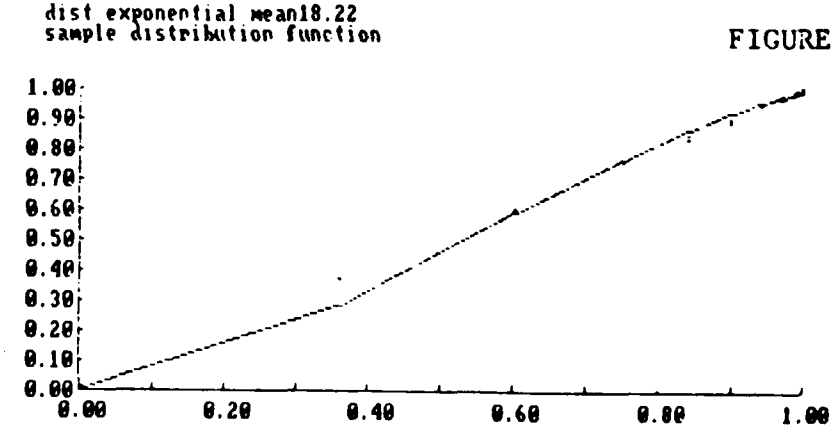


FIGURE 6

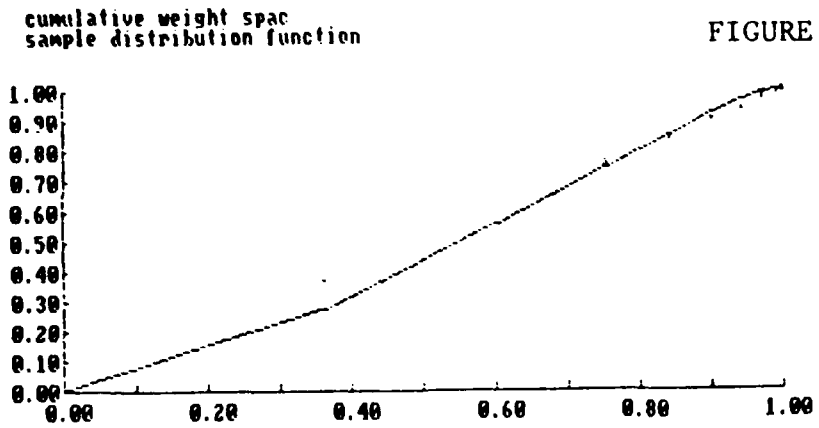


FIGURE 7

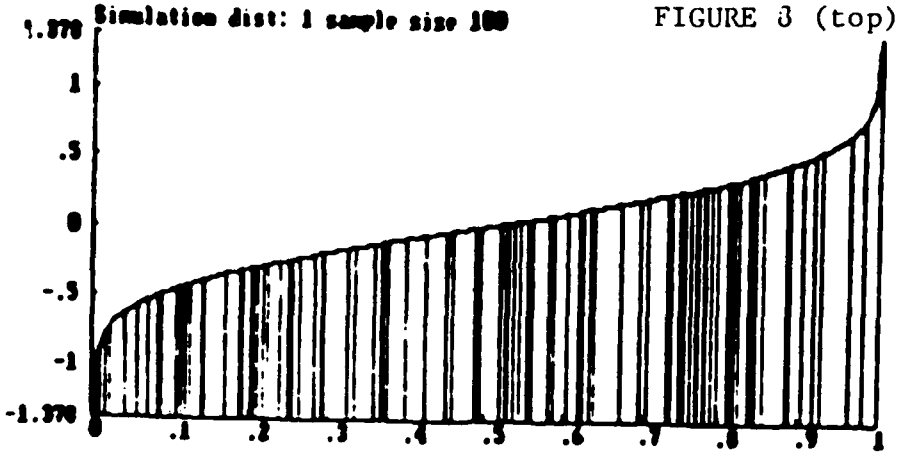


FIGURE 3 (top)

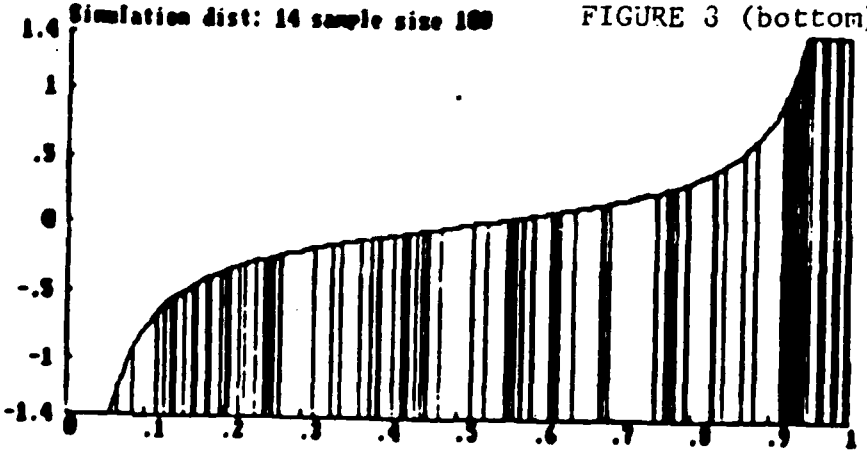


FIGURE 3 (bottom)

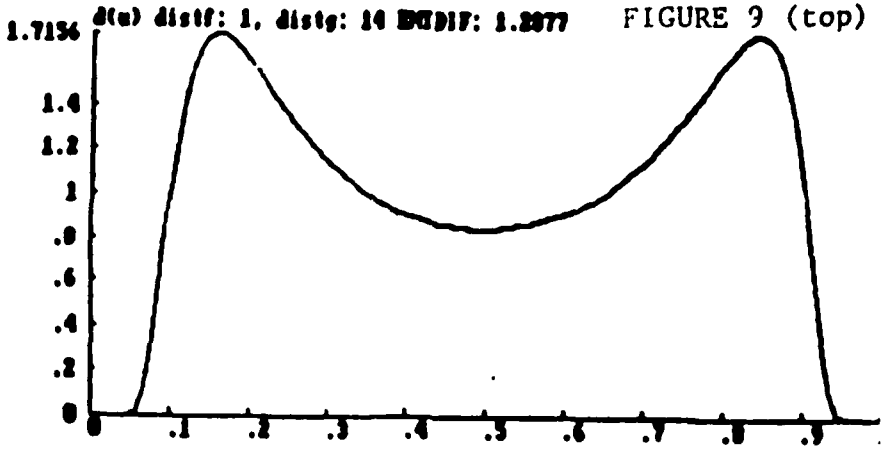


FIGURE 9 (top)

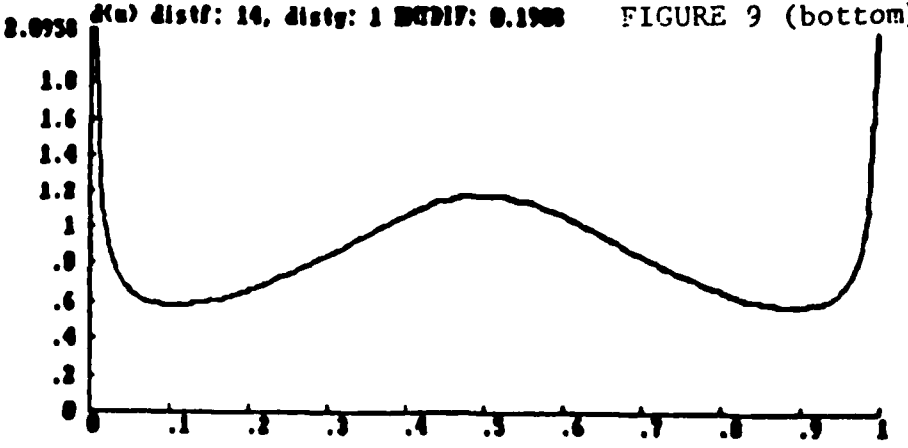


FIGURE 9 (bottom)

END

11-87

DTIC